informationresearch

VOL. 16 NO. 4, DECEMBER, 2011

Authoring Wikipedia articles as an information literacy assignment: copy-pasting or expressing new understanding in one's own words?

Eero Sormunen and Leeni Lehtiö
School of Information Sciences, 33014 University of Tampere, Finland

Abstract

Introduction. We report on a pilot study of students' use of sources in authoring Wikipedia articles. The procedure is demonstrated by investigating how students processed texts from sources to compose their own texts.

Method. Four groups of upper secondary school students (aged 17-18 years) participated in an eight-week geography course and seven groups in a corresponding biology course. Each group wrote an article for Wikipedia as part of the course programme. The students' work was observed and they were interviewed. Published Wikipedia articles, cited sources and sources identified in a plagiarism test were also included in the dataset.

Analysis. Article sentences and sources were compared and text transformations were classified using an ordered five-category typology ranging from 'copy-paste' to 'synthesizing across sources'. Descriptive statistics were used. Observation memos and interview transcripts were analysed qualitatively.

Results. Students used almost solely Web-based sources. The researchers estimated that about 30 percent of sources used were not cited. The analysis of cited and other identified sources showed that about one third of Wikipedia article sentences were verbatim or slightly edited copies of source sentences.

Conclusions. A more analytical approach is needed in the study of the core processes of source-based writing: making meaning *from* sources and making meaning *for* texts written.

Introduction

Writing is a traditional method used in schools to enhance learning by making students process information and construct knowledge (see e.g., Tynjälä *et al.* Lonka 2001). Writing in a particular genre merges the authentic practices and norms of that genre into the learning process and gives a framework for the task. In a science class it is natural to write in the form of a research report and to emulate good practices of scientists (See e.g., Li and Lim 2008, Chu *et al.* 2008). In information literacy instruction writing based on sources searched by students themselves is a widely used type of learning assignment. We call this assignment a *source-based writing* task.

Source-based writing assignments may take different forms (e.g., writing a review of a topic, solving a problem) depending on the framing genre, the instructional design of the task and the specific goals given by the teacher. However, the core of the task is that students search and study multiple texts and compile another text. The aim is that students read sources with thought, construct knowledge on the given topic, and based on that knowledge and available sources compile a new text indicating what they have constructed and learned. In information literacy instruction we assume that through this practical exercise of source-based writing students learn, along with subject contents, to search for, evaluate and use information effectively and ethically. [Note: In this paper, we use the terms *knowledge* and *information* as defined in the cognitive viewpoint (see Ingwersen and Järvelin 2005) and the terms *knowledge construction* and *learning* as defined in the tradition of (social) constructivism (see Spivey 1997).]

In the Internet age, copy-pasting has become a widely recognized problem that makes it difficult to achieve the intended learning objectives in source-based writing tasks. An obvious risk is that students will transfer information mechanically from the sources into their own texts instead of transforming it in the cognitive process of knowledge construction. In an unfortunate case, the student fails to achieve learning goals in topical contents as well as in information literacies. Surprisingly, copy-pasting has rarely been studied in the information literacy instruction context. Even then the primary focus has been on plagiarism (see McGregor and Streitenberger 2004; McGregor and Williamson 2005; Williamson and McGregor 2006; Williamson *et al.* 2007; one exception: Alexandersson and Limberg 2003).

Copy-pasting is regarded here as a situation where a piece of text from a source document is used verbatim in one's own text (transported as expressed by Alexandersson and Limberg 2003). Copy-pasting becomes plagiarism if the source document is not duly acknowledged. From the knowledge construction viewpoint, the primary problem in copy-pasting is that the learner is not processing information in order to learn about the topic. In terms of information literacy, the copy-pasting person is not willing to learn effective and ethical practices of information assessment and use. Plagiarism overlaps with copy-pasting but has aspects such as thefts of ideas which do not appear as verbatim copying (see McGregor and Williamson 2005). In this paper we limit ourselves to the copy-pasting problem but take advantage of plagiarism research when appropriate.

Our aim was to study copy-pasting behaviour using Wikipedia as the framework of source-based writing. We consider that the Wikipedia genre offers an interesting opportunity to enhance traditional information literacy instruction models in solving the copy-pasting problem. Students are familiar with Wikipedia as an information source (e.g., Head and Eisenberg 2010), Wikipedia requires consistent use and citing of sources (Purdy 2009, 2010) and the articles are published making the task authentic and meaningful (Forte and Bruckman 2010). All these attributes of Wikipedia have the potential to persuade students to assess and process information for constructing knowledge (see the next section for more detailed justifications).

We conducted a pilot study on a Wikipedia authoring assignment in an upper secondary school. The overall goal of the pilot was to learn how Wikipedia works as the framework of source-based writing assignments in information literacy instruction. In this paper we report findings related to students' use of information sources in authoring Wikipedia articles. We scrutinized the phenomenon at micro level: how much do the students process texts available in sources beyond copy-pasting in producing their own texts.

The paper is organized as follows: first, a short review is given on related literature. The goal is to build a framework to understand copy-pasting as a learning problem and how information use might be studied in source-based writing assignments. Secondly, we present our research questions and the methods of data collection and analysis. Then the main findings are reported. Finally, we discuss the findings and present some conclusions.

Related research

Learning orientation and copy-pasting

Past research suggests that the risk of copy-pasting is associated with the student's surface orientation in learning. Limberg (1999) observed that low-performing students had a tendency to "fact finding" approaches in independent learning tasks which easily lead to copy-pasting. These students are not interested in the genuine inquiry of the topic. They see the assignment as a search exercise where "right answers" are collected from sources and transferred into the research paper (Alexandersson and Limberg 2003). Limberg and her associates (2008) summarize the findings of several empirical studies in Swedish schools and emphasize that motivation to work on sources is connected to interest in learning about the topic of the assignment.

Heinström's (2002, 2006) findings concur with those of Limberg. She found that students with a high level of engagement in the topic of the assignment tended to adopt strategic or deep learning orientations. Strategic and deep learners were portrayed as deep divers as searchers: they invested considerable effort in seeking high quality sources and in analysing their content. Heinström (2006) equates the fact finding approach with fast surfing behaviour in information searching. She argues that fast surfers simplify information literacy as searching skills and fail to practice assessment and use of acquired documents.

Plagiarism studies

Researchers of plagiarism in school assignments have mainly applied ethnographic methods to scrutinise the phenomenon and extended the view by quantitative analysis of plagiarism and copying in students' research papers. McGregor and Streitenberger (2004) observed that levels of copying and plagiarism were higher among those students who concentrated more on the format of the end product (*looking good*) than on the process of gathering and synthesizing information for the sake of its content. The authors conclude that process-oriented students tended to manipulate information more deeply, internalize their topics, and maintain an interest in their topic. In a later

study, McGregor and Williamson (2005) found that plagiarizing students were less engaged with their topics, less focused on the learning process and remembered less about the topic in a later test. The authors concluded that less plagiarizing students tried more to make sense, seek meaning, think, do research and learn with the help of information sources.

McGregor and Streitenberger (2004) measured plagiarism scores for research papers written in two English classes. The authors define the high score for plagiarism: 30 percent or more of the text was almost verbatim copy from sources. The percentage of papers where plagiarism scores were fairly high varied between 10 and 31 percent. The lower percentage of plagiarism in one class was attributed to the many reminders by the teacher to avoid plagiarism. She actively instructed students how to quote and cite appropriately. However, the analysis of texts revealed that the teacher's interventions to reduce plagiarism did not lead to more intensive processing of information. Plagiarism did not change into proper paraphrasing and acknowledging of sources but into copying and careless citing. The authors call these students' behaviour scribing, which appears in three different forms: legitimate copying (quoting), inappropriate copying (copy-pasting or near copy-pasting) and plagiarism.

McGregor and Streitenberger (2004) developed a five-level classification for the comparison of texts in student reports and sources used. It was also applied in the later study by McGregor and Williamson (2005). The categories of copying were defined as follows:

A. No copying.
B. Paraphrasing, doesn't closely resemble original.
C. Paraphrasing, can easily recognize original pattern of sentences and paragraphs, but many words have been changed.
D. Copying, with phrases rearranged, omitted, some words added. Occasional synonyms used.
E. Copied word-for-word for the most part. May involve some omissions, slight rearranging, minimal changing of tenses, minimal use of synonyms.

In the analysis, the authors calculated how large a percentage of the report's text was copying sources at levels D and E. The students were allocated to the categories of least and most plagiarizing writers depending on D&E percentages. However, the unit of analysis was not explicitly specified. The examples indicate that it ranged from one to several sentences, and even across text paragraphs.

Reading-to-write studies

A special community of scholars in education is interested in *reading-to-write tasks*. The task of reading multiple texts and composing a new one is called *discourse synthesis* or *writing from sources* (see Spivey 1997; Segev-Miller 2004). The focus of reading-to-write studies is on how a reader presents in written form what he or she has learned from one or more texts (McGinley 1992; Boscolo *et al.* 2011). In this research tradition, research subjects (students) are typically given two or more source texts and specific instructions to write their own text. Researchers collect and analyse data on the process and resulting texts.

In her seminal book Spivey (1997: 136) defines the reading-to-write task as an act where a person is concurrently in two roles: in the role of reader building meaning *from* a text and in the role of writer building meaning *for* a text. Reading (comprehension) and writing (composing) are not linear,

consecutive processes but tend to blend. The writer reads others' texts (sources) but also his or her own text when composing it. Writing starts as a cognitive process while reading in the form of planning how the sources can be used in the text to be written. In addition to sources, the writer takes advantage of his or her knowledge of the topic and understanding of the discursive practices of the intended audience (Spivey 1997: 144-145).

Empirical studies on reading-to-write tasks show that making a synthesis of multiple texts is cognitively more demanding than writing a summary of a single text (Davis-Lenski and Johns 1997; Mateos and Solé 2009). In *summarizing* a single text it is possible to maintain the structure of the original text. The *synthesis* of multiple texts requires an integrating idea (*superproposition*) of how to transform information from differently structured, even contradictory, texts into a new structure. The synthesis requires knowledge transformation to a greater extent than making a summary (Mateos and Solé 2009; Segev-Miller 2004).

Spivey (1997: 149-163) gives a description of methods used in analysing the relationships between sources and texts written by students. She used discourse analysis to parse source texts (in this case three) and texts written by students into propositions called content units, e.g., {benefit, armadillo, mankind}. Content units were further organized into a hierarchical structure. Based on this semantic framework, the researcher could analyse, for example, the unique and overlapping contents of source texts, what contents students had selected from sources, how they had structured their text and what type of connections (synthesis) they had made to render it integrated and consistent.

Wikipedia as a writing framework

Wikipedia is a participatory encyclopaedia built on a wiki (for the English version visit: http://en.wikipedia.org/wiki/Main_Page). Anybody is invited to contribute, even anonymously, by writing and editing articles as one of the Wikipedians. Its authoring policy rests on three principles: (1) *verifiability*, (2) *no original research* and (3) *neutral point of view* (Huvila 2010; Sundin 2011). All facts in articles should be verified by references to reliable published sources external to Wikipedia. Articles should not contain new information or interpretations which have not been published earlier in some trusted forum. The third principle reminds writers to present competing views or balance between them in contradictory issues. For a more detailed discussion, see Huvila (2010); Lyut and Tan (2010) and Sundin (2011).

The findings of past research highlight some aspects of Wikipedia which make it an authentic and meaningful framework for source-based writing assignments:

1. Students are familiar with the genre of Wikipedia articles because they exploit it regularly for various information needs (e.g., Head and Eisenberg 2010, college students; Harouni 2009, high school students). Most students have at least some personal experience of Wikipedia; a good starting point for knowledge construction and engagement.
2. Harouni (2009) observed that high school students often use Wikipedia uncritically. Lim (2009) reports that college students seem to trust Wikipedia because their everyday experiences are primarily positive. On the other hand, Lim (2009) also reports that students are confused since they are aware of quality problems in Wikipedia. This tension calls for instructional interventions that help students to scrutinize the problems of information evaluation – a learning goal serving their personal needs.

3. Sundin (2011) made an ethnographic study of everyday practices of Wikipedia editors and found that construction of knowledge and referencing external sources are quite transparent processes. He concludes that this makes Wikipedia an excellent forum to discuss and demonstrate the credibility of information. This is a core issue in information literacy instruction.
4. The studies by Forte and Bruckman (2007; 2010) show that writing an article on a public wiki engages students deeply in learning the subject contents, in assessing the reliability of sources, in citing them and in trying to write a *high quality* article. The engagement is primarily based on students' perception that they are writing for a real audience whose expectations they do not want to disappoint.
5. Purdy (2009; 2010) argues that writing and citing guidelines for Wikipedia are similar to the requirements of scientific writing (cf. Sundin 2011). Jennings (2009) compared Wikipedia guidelines with the information literacy standard of the Association of College and Research Libraries (2000) and found that they overlap quite a lot. Basically, Wikipedia offers a framework for information literacy instruction which is similar to that of a scientific article and concurrent with the information literacy standards.

Students face quite a different discourse in the traditional information literacy instruction (see e.g., Achterman 2005; Julien and Baker 2009). In the traditional setting, students are put to work on an *imposed* assignment (cf. Gross 2005) which is seldom contextualized into students' reality.

Comments on related research

Learning orientation and copying. The development of the information search process model and the extensive research associated with it give us a solid basis to study learning tasks as construction processes (see Kuhlthau 2004). The studies by Limberg (1999) and Heinström (2002) reveal the connection between the student's engagement (motivation), learning orientation and use of sources for knowledge construction. However, this line of research has not focused on the actual use of sources: how students transfer or transform information from sources to their own text.

Research on plagiarism. Studies on plagiarism in school assignments have mainly used ethnographic methods to describe the phenomenon of plagiarism. A side track of research the quantitative analysis of written texts and sources used, has contributed with a simple categorization schema for calculating the degree of copying in texts (see McGregor and Streitenberger 2004; McGregor and Williamson 2005). The work gives us a valuable model of how to compare the texts of written reports and sources used. The weakness of the analysis conducted so far is that the unit of analysis is defined ambiguously (not explicitly fixed to the text's phrase, sentence or paragraph structure) and the definition of copying categories is quite shallow.

Reading-to-write studies. It is surprising that research in information literacy instruction has totally ignored the work in reading-to-write studies. This line of research is really focused in the core process of source-based writing tasks and can help in developing more powerful conceptual frameworks for studying source-based writing phenomena. The scholars in the field have adopted an experimental research strategy and developed semantic analysis procedures to represent text contents and transformations between texts. This is an appealing approach to study how writers process and synthesize source texts when composing a new text. However, the analysis of texts at the level of propositions requires a lot of resources as the number of texts increases. The number of sources cannot be limited in realistic source-based writing tasks, and further, *n* students typically

write on *n* topics. However, the methods used in reading-to-write studies point in a direction where one could find relevant methods.

Wikipedia as a source-based writing framework. The studies reported above reveal the fundamental challenge of the information literacy instruction: how to make source-based writing assignments meaningful to students so that they engage in constructing knowledge on the topic of the assignment. Wikipedia seems to possess motivational benefits which are difficult to achieve in the traditional setting based on the research paper metaphor. In this paper, Wikipedia is not in our primary interest but gives us an interesting framework to study how students use sources in source-based writing assignments.

Research questions

The objective of our pilot project was to explore how writing for Wikipedia works as an information literacy exercise and as a group assignment in an upper secondary school. In this paper we report the findings related to the following research questions:

1. What type of information sources do students use while authoring Wikipedia articles?
2. To what extent do students copy or process information from sources for the Wikipedia article?

Answering the first research question builds a base for answering the second research question. By *use* we refer here both to cited and non-cited (plagiarized) information sources. In a pilot study we are not only interested of empirical findings but try to construct an overall picture of the phenomenon and learn how to study it.

Data collection and analysis

Case courses

Data were collected from two eight-week courses in an upper secondary school in the city of Tampere, Finland, during the spring term 2010. The original idea of the course came from teachers and they had a free hand to design the courses. Ten students took the first course in geography and sixteen students took the second one in biology. The students were organized into eleven project groups. Each group wrote one new article or made major extensions to an existing article in the Wikipedia Finnish edition. Two teachers were involved in the process, each in charge of one course.

Both teachers prepared a list of possible article topics for each student group to select one. During the first session the teacher introduced the Wikipedia assignment. The librarian working for the development project of information literacy instruction introduced the basics of Wikipedia, the requirements for the Wikipedia articles and gave a demonstration of some online reference sources. Some instruction in using and citing sources was also given in spoken and printed forms. Student groups searched for information and wrote their own texts during weekly sessions in the computer classroom. They worked at their own pace and the teacher was available for help, discussions and

encouragement. The teacher also tried to control how each group made progress and reminded them of deadlines.

Students were instructed to complete their texts offline on their computers, give the file to the teacher for assessment and after her acceptance publish it in Wikipedia. The teachers wanted to control the quality of the contributions. They also considered the assessment of students' articles easier when they could be sure that the texts were not edited before their assessments by someone in the Wikipedia community.

In total, the students of the geography (biology) course worked on the article for nine (eight) 45-minute sessions in the computer classroom over a period of eight (seven) weeks.

Data collection

We collected a rich dataset from the process. In the geography course, we gathered basic data on students with a pre-questionnaire, observed all weekly meetings in the computer classroom, and interviewed the project groups at the end of the course. The teacher was interviewed both before and after her course. We also collected all instructions and of course the articles uploaded into Wikipedia. In the biology course, we replaced observation by contextual inquiry interviews (see Raven and Flanders 1996) because of technical and methodological problems in observations. The noise in and around the classroom made it difficult to hear the discussions. The observer obviously had a disruptive effect. The students tended to fall silent when the observer turned her attention to their group.

The findings reported in this paper are based on the content analysis of the Wikipedia articles written and sources used in writing them. Observations and interviews have a complementary role. Our first task was to identify not only the cited sources but the unacknowledged (not cited) sources. From the interviews with the students we knew that none of them had visited the library but used the Web as the primary information channel. This made the hunt for the non-cited sources a manageable problem.

We approached the problem by dividing each Wikipedia article into sentences. Next, the sources used in writing each sentence were hunted for through the following procedure:

1. The sources cited at the paragraph level were first checked.
2. If a sentence did not match any source cited within the text paragraph, other sources listed in the reference list were checked.
3. If a sentence could not be linked to any of the sources cited, a plagiarism checking procedure was applied. The first step was to conduct test searches with Google using up to six meaningful words from each sentence. We checked the sources on the three first pages of search results to find texts matching the students' sentences. In case of a match, we checked that the sources found had not been updated after the students' article had been published.
4. As a final attempt we searched the online encyclopaedia available in the school computer network. We also browsed relevant parts of printed articles and books which the teacher had brought into the classroom.

In the identification of sources we compared the literal and semantic content of each article sentence with the content of candidate sources (including tables and images). If similar information was found in several sources, we tried to find out if the sentence could have been formulated using only one source. If that was not the case, we checked if the sentence could be composed by combining information from two or more sources.

Data analysis

In the analysis of text transformations we were not completely satisfied with the plagiarism categories and the units of analysis used by McGregor and Streitenberger (2004). The categories at the '*No copying*' end of the scale especially were difficult to make operational. We found it very difficult see the operational difference between *paraphrasing* and *no copying*. If the task is to write a text based on information interpreted from a source, when do we know that the threshold of *no copying* has been exceeded?

We also decided to use a sentence as the unit of analysis instead of flexibly selected text extracts to improve the repeatability and reliability of the analysis. In the course of analysis it is possible expand the text window to explore larger units of text if appropriate.

Using the categorizations found in the literature (See Spivey 1997; Davis-Lenski and Johns 1997; Mateos and Solé 2009) as a guiding frame we derived inductively from our data five categories to describe the degree and nature of information transformation from the source to the article text:

1. **Copy-pasting.** A verbatim copy of a source sentence. Appropriate if cited and formatted as a quotation.
2. **Near copy-pasting.** A mechanically edited sentence from the source (e.g., syntax changed, words of minor importance removed).
3. **Paraphrasing.** Expressing the key information of a source sentence in the writer's own words.
4. **Summarizing from a single source.** Writing a sentence in the writer's own words based on two or more sentences in a single source.
5. **Synthesizing across sources.** Writing a sentence in the writer's own words based on sentences in two or more sources.

The categories are sorted in order of increasing transformation of information from verbatim copying via paraphrasing to synthesis of information. The aim of specifying categories the way they are was to make the boundaries relatively easy to distinguish in the analysis.

The use of categorizations may be sensitive to varying interpretations by different analysts. However, in a pilot study we consider it adequate to use straightforward procedures. The second author made the identification of potential sources alone. The classification of source-article transformations was then made independently by both authors. The consistency between the two analysts was 80.0 percent. The inconsistencies were discussed to reach a consensus. The analysis was conducted on 233 sentences in eleven articles where at least one of the sources used could be identified. We could not map twelve sentences into any source candidate and they were excluded from the analysis.

Table 1 summarizes the basic data of the Wikipedia articles composed. The teachers did not set definite requirements for the length of articles or the number of sources. Thus the length of the articles ranged from 5 to 57 sentences and the number of sources from 2 to 12. In four groups (B, F, J and L) we could identify 2-4 sources which the students had not cited (from 25 to 40% of all sources). In five articles we did not identify missing references.

Class	Group	Topic of the article	No of words	No of sentences	No of cited sources		No of non-cited sources	No of known sources	No of sentences per source
					in Fin	in Eng			
Geology	A	Desertification in Africa	651	57	11	0	1	12	4.7
	B	Prevention of desertification	429	33	6	0	3	9	3.7
	C	Wind erosion	83	5	2	1	0	3	1.7
	D	Water pollution	500	30	2	4	1	7	4.3
Biology	F	Sulphur dioxide	139	13	4	0	4	8	1.6
	G	Lead	233	17	1	6	0	7	2.4
	H	Tropospheric ozone	151	14	2	1	0	3	4.7
	I	Hydrogen sulphide	84	8	1	1	0	2	4.0
	J	Mechanical particles	323	28	6	0	2	8	3.5
	K	Nitrogen oxides	156	12	6	0	0	6	2.0
	L	Hydrocarbons	332	28	3	0	2	5	5.6
Total			3081	245	44	13	13	70	3.5
Average			275	22	4	1	1	7	3.5
Median			233	17	3	0	1	7	2.4

Table 1: Basic attributes of authored Wikipedia articles.

Findings

Information sources cited and not-cited

The analysis of sources used (n=70) revealed that students really preferred Web sources. Only three groups used printed sources (two encyclopaedias or one geography textbook). In the Web students trusted public corporate sources (see Table 2). The share of sites of public administration at national and regional levels (governmental sites) and sites of associations and non-profit public

organisations (communal sites) was about 53 percent in the sources used. Reference sources including Wikipedia and textbooks accounted for 19 percent. One tenth of sources (10%) were teaching materials such as presentation slides and web teaching materials. Other sources had a minor role: Web sources published by mass media companies (newspapers, broadcasters) 7 percent, students' papers and pages on commercial sites 4 percent each. Two web sources could not be opened and were left uncategorized.

Group	Communal sites	Governmental sites	Reference sources	Wikipedia	Media	Commercial sites	Teaching materials	Student papers	Dead links	Total	%
A	3	1	2	3	1	0	1	1	0	12	17 %
B	1	1	2	1	1	0	2	1	0	9	13 %
C	1	0	1	0	0	0	0	1	0	3	4%
D	3	1	0	1	1	1	0	0	0	7	10 %
F	2	4	1	0	0	0	1	0	0	8	11 %
G	0	3	0	1	0	2	0	0	1	7	10 %
H	1	2	0	0	0	0	0	0	0	3	4%
I	1	1	0	0	0	0	0	0	0	2	3%
J	6	0	0	0	2	0	0	0	0	8	11 %
K	1	2	1	0	0	0	1	0	1	6	9%
L	0	3	0	0	0	0	2	0	0	5	7%
Total	19	18	7	6	5	3	7	3	2	70	100 %
%	27%	26%	10%	9%	7%	4%	10%	4%	3%	100 %	
Non-cited	3	3	0	3	1	0	2	1	0	13	

Table 2: Types of information sources used by different groups. (Thirteen non-cited sources are included in n= 70 and also presented separately in the last row.)

Our plagiarism test revealed that thirteen sources (about 19% of all sources) we assume the students had used were not mentioned in reference lists. Most non-cited sources belonged to categories communal sites (3), governmental sites (3), Wikipedia (3) and teaching materials (2). However, the relative risk of being plagiarized was highest for Wikipedia articles, student papers and teaching materials. The sum of non-cited Wikipedia articles, student papers, and teaching materials together was 6 out of 16 (31 %). For other source types the rates of missing citations were lower: communal sites 3/19 -> 16%, governmental sites 3/18 -> 17%, the rest 1/15 -> 7.5%. The use of student

authored web pages and Wikipedia articles were against the explicit instructions of the teacher. This may have affected students' tendency to concealing some of the sources used.

Copy-pasting and text transformations

In the interviews students told how they had used sources. They were aware that they were not allowed to copy-paste. Some groups first compiled notes and wrote their article based on their notes. Other groups mentioned that they read sources first and then tried to rewrite the text in their own words. Most groups wrote the plain text first and attached references to them afterwards. Only one group confessed that they had copied one sentence because it was so well constructed.

The analysis of articles showed that about five percent of article sentences were exact copies of source sentences (see Table 3). Another thirty percent of sentences had been edited slightly but in a mechanical way. Thus, in one third of sentences we could not see any sign of the writers' cognitive effort in transforming information. The good news is that two thirds of texts were obviously processed beyond copy-pasting.

Group	Percentage of sentences					Number of sentences analysed	Source unknown	Total number of sentences
	Copy-paste	Near copy-paste	Paraphrasing	Summarizing	Synthesizing			
A	4%	43%	32%	7%	14%	56	1	57
B*	15%	19%	30%	26%	11%	27	6	33
C	0%	20%	40%	20%	20%	5	0	5
D	0%	59%	31%	7%	3%	29	1	30
F	0%	54%	38%	8%	0%	13	0	13
G*	0%	47%	18%	18%	18%	17	0	17
H	0%	14%	43%	7%	36%	14	0	14
I*	0%	0%	38%	25%	38%	8	0	8
J*	8%	12%	42%	27%	12%	26	2	28
K	10%	20%	30%	10%	30%	10	2	12
L	7%	4%	54%	21%	14%	28	0	28
Average/Sum	5%	30%	36%	15%	15%	233	12	245

Table 3: Degree of information processing in writing Wikipedia articles: matching article sentences to sources used.
*) Because of rounding-up, the sum of the category percentages is 101%.

Paraphrasing is the most common way of using source texts (36%). The writer expresses the contents of a single sentence in her/his own words. The semantic elements of the sentence have

been modified to serve the goals of the writer better, for example, to clarify the message, to modify phrases to fit the sentence into the neighbouring text or to shorten the text. The limitation of paraphrasing is that the writer's view might focus on a single sentence at a time. Summarizing (15%) indicates that the writer tries to condense the message of two or more sentences of a source into one sentence while in synthesizing (15%) the message is condensed across two or more sources.

A look at individual articles reveals differences between groups. In groups A, D, F and G a large share of sentences (47 - 59%) were verbatim or near copy-pastes indicating a low level of engagement in processing information. Interestingly, none of sentences in groups D, F and G were exact copy-pastes. Had these students adopted a conscious strategy to minimise their effort in meeting the *do not copy-paste* requirement? On the other hand, the articles by groups A and G contained many summarised and synthesized sentences (21% and 36%, respectively) while groups D and F seldom processed information beyond paraphrasing (summ+synth 10% and 8% respectively).

Paraphrasing source texts (as operationalized in our study) indicates that the writers' information processing goes beyond copy-pasting. However, paraphrasing, like copy-pasting allows a sentence-by-sentence reading-to-write approach. In the above-mentioned groups D and F, at least 90 percent of article sentences were near copy-pasted or paraphrased. While the writers were able to avoid direct copy-pastes they demonstrated quite an elementary sentence-by-sentence knowledge construction practice. Most source sentences were edited mechanically and some of them were paraphrased during the writing process.

Many groups showed a tendency to write many paraphrased sentences but also to go beyond that approach. For example, groups C, H, J and L paraphrased a lot (40% or more of texts) and in addition, they actively summarized within sources and synthesized across sources (the sum varied between 35% and 43%). Group I achieved the highest percentage of summarized and synthesized sentences (63%) but their article was short (eight sentences). These groups demonstrated a reading-to-write practice which is not limited to the sentence-by-sentence approach. The writer processes larger chunks of text within a source or across sources requiring more transformations in composing a coherent text of his or her own.

The above examples characterized articles where either low-end or high-end text transformations dominated. The articles by groups B and K are different from these since they contain both a high share of copy-pasted sentences (10 and 15%) and summarized and synthesized sentences (37 and 40%). The distribution of sentences is quite even across all categories. A likely explanation for the diversity of writing behaviour is that some groups were not engaged in collaborative authoring. According to the interviews both groups B and K divided the writing task early in the process and merged individually compiled texts at the end.

The length of the articles varied from five to fifty-seven sentences. This makes the use of percentages in the comparison of groups problematic. For example, the weight of a single sentence is twenty percent in the shortest article (length five sentences) while only less than two percent in the longest one (length fifty-seven sentences). The articles by groups A, B, J, and L contained the highest number of summarizing and synthesizing sentences (12, 10, 10, and 10 respectively). The article by group A was exceptionally long and dominated by low-end transformations (two verbatim and twenty-four near copies). In groups J and L the orientation was towards high-end transformations. The article by group B was an example of evenly distributed transformations.

We could not map twelve out of 245 sentences on any cited or non-cited source (Table 3, second last column). This small set of sentences looked similar to others in terms of expressing facts that should be verified by proper sources. We could be quite sure that they were not based on sources cited by the authors. The plagiarism procedure did not succeed in identifying sources. However, we assume that the exclusion of twelve sentences (about 5% of all) hardly had a substantial effect on our empirical findings.

Discussion

As could be expected, the students preferred to use Web sources. In the interviews, the students reported that their textbooks contained too general information to be useful for writing about the specific topics of the assignments. Only in a few cases they used the printed materials made available in the classroom by one of the teachers. The librarian who gave the introductory demonstrations of information resources advocated library services but none of the students visited the school library or the nearby public library.

The finding that the students used a lot of materials from communal (especially in the environmental geography class) and governmental sites (especially in the biology/hazardous substances class) may be related to the themes of the two courses. This observation has a connection to an even more interesting detail: None of the groups used scholarly publications although a lot of such materials are available in the Web. Forte and Bruckman (2010) found that high school students working on a research paper used Wikipedia and other Web sources to comprehend their topic but cited scholarly sources because they had to. Genuine scholarly publications were simply perceived to be too difficult in the construction of knowledge. The students also felt that scholarly texts are so complex that it is difficult to transform their content into a text written for a layman audience (of a public wiki). Web resources such as governmental and communal sites are often designed to serve a general audience or heterogeneous groups of professionals. The genre of publications makes a difference for a student author.

Our plagiarism procedure revealed that one fifth of the sources used were not included in the reference lists. Further, we found that about five percent of sentences in the articles were obviously drawn from sources we could not identify. If these figures are combined we may estimate that in this student group plagiarism in using sources may rise to about thirty percent (twenty-five of eighty-two sources not cited). We do not know if the figures should be regarded as high or low because we lack relevant comparative data from earlier studies. The rates of plagiarism reported by McGregor and Streitenberger (2004), and by McGregor and Williamson (2005) are not easy to interpret and not directly comparable to our results.

Nilsson et al. (2005) argue that copy-pasting and plagiarism can be approached from two perspectives: students are positioned (1) as cheaters raising an ethical problem or (2) as learners raising a pedagogical problem. In light of the interviews, the problem seems to be more pedagogical than ethical. Our students had not mastered the basic routines of source-based writing. They often copied extracts from sources into their computers or composed personal notes when reading sources and started writing the article without keeping a record of sources read. Several students explained that they marked in-text citations and composed the lists of references after the article text was completed. Not all sources could be recalled and a lot of errors were made in attributing in-text citations. Similarly to the case reported by McGregor and Streitenberger (2004) the teachers did not instruct their students in concrete terms how to work with sources. In this situation students, as also

reported by Forte and Bruckman (2010), perceived citing as an extra effort added on to the primary task of writing. In genre terms: in typical school texts citing of sources is not an issue. Students are required to focus only on arguments and language.

The comparison of article sentences and their sources revealed that verbatim copy-pasting was not very common but some student groups had an obvious tendency for near copy-paste behaviour. Students avoided the use of direct copy-pasting but instead did small mechanical editing. The finding suggests that these students adopted a surface orientation to learning and were not engaged or had not mastered source-based writing practices (cf. Heinström 2002; 2006). However, in most groups the main parts of texts were based on paraphrasing, summarising within sources and synthesizing across sources.

The findings suggest that most students were engaged in working with source texts. In the best case this means that students (1) constructed meaning from texts to be able (2) to construct meaning for the text to be written. This is called discourse synthesis by Spivey (1997: 146). From the tradition of discourse synthesis studies we borrowed two text transformation concepts: summarizing and synthesizing. These two concepts helped us to make the higher level categories of text transformation operational and exceed the analytical limits of categorizations introduced in the study of plagiarism (cf. McGregor and Streitenberger 2004).

Although students paraphrased, summarized and synthesized source texts this does not guarantee that they exceeded the fact-finding approach (see Limberg 1999). From the information literacy instruction point of view, Wikipedia articles can be regarded as a structured collection of verified facts (typical of the encyclopaedic genre). The genre, when applied, has some pedagogical implications. Wikipedia obviously supports learning goals related to source-based writing but may be less suitable for teaching higher level information literacies and creative academic skills.

Wikipedia assignments cannot be expanded, for example, to the analysis of personal empirical data or the study of speculative/imaginary research ideas. Thus, genuine inquiry processes are not possible in this genre. On the other hand, our results show that even in this context it is possible to practise discourse synthesis. The student has to make meaning from texts in different genres and transform texts to be able to make meaning for the text written in the Wikipedia genre. The experiments conducted in reading-to-write studies (see Spivey 1997) demonstrate how complex a task is the comprehension of texts in various genres and from various contexts.

One limitation of the pilot study was that the use of source texts was analysed at the sentence level, only. Although the basic concepts of information transformations are convenient to operationalize using sentences as the units of analysis, it is important to cumulate data at higher level text structures to make findings more meaningful and applicable. Further, the findings call for future research on how the advanced use of sources is related to

- the number and complexity of sources used;
- knowledge construction (e.g., quality of articles);
- learning experiences;
- personal characteristics of learners (e.g., personality, learning orientation);
- group behaviour (e.g., activity and ways of collaborating);
- instructional design (e.g., teacher involvement).

Conclusions

Our goal in the pilot study was to find out how students use information in writing Wikipedia articles as a school assignment, and specifically, how they process information in source-based writing. We introduced a procedure to analyse how the texts read as sources were used in the writing of the text for Wikipedia articles. The empirical findings suggest that students use copy-pasting but also transform texts by paraphrasing, summarizing within sources and synthesizing across sources. We could demonstrate that our procedure could be used to reveal different types of text transformations.

Our study also identified a neighbouring research community, reading-to-write studies, which could obviously contribute to research on information literacy instruction and especially in research on source-based writing assignments. The conceptual frameworks, methods and empirical findings of this community could help in focusing research on specified information literacy assignments. Past research on information literacy instruction has prioritized assignments simulating genuine inquiry and writing in the genre of a scientific paper. This is a reasonable approach but not sufficient since researchers have been blind to (1) information environments and publishing genres other than those of science (for exceptions see Forte and Bruckman 2010) and (2) the most elementary practices of learner-source interaction (i.e., the act of discourse synthesis).

The pilot study demonstrated that the procedure for the analysis of text transformations provides us with a useful tool in studying students conducting source-based writing tasks. The procedure needs to be further developed and tested to become a validated method. The next step is to apply it to a larger dataset (already collected at the time of writing).

Acknowledgements

The study was part of the Know-Id project and the first author's sabbatical project funded by the Academy of Finland (grants no. 132341 and no. 136401). The authors thank the teachers of the case courses and the *Tieto haltuun* project in the City of Tampere for cooperation in data collection. We are grateful to Jannica Heinström, Carol Kuhlthau, Ross Todd, colleagues at the University of Tampere, and two anonymous reviewers for constructive comments to improve the manuscript.

About the authors

Eero Sormunen is a Professor in the School of Information Sciences, University of Tampere, Finland. Sormunen received his Master of Science (Electrical Engineering) in 1978 from the Tampere University of Technology and his PhD (Information Studies) in 2000 from the University of Tampere, Finland. He can be contacted at: eero.sormunen@uta.fi.
Leeni Lehtiö received her Master of Science (Information Studies) in 2011 from the University of Tampere, Finland. Currently she works as an Information Specialist in the University of Turku Library.

References

- Association of College and Research Libraries. (2000). *Information literacy competency standards for higher education*. Association of College and Research Libraries. Retrieved 9 May 2011 from http://www.ala.org/ala/mgrps/divs/acrl/standards/informationliteracycompetency.cfm. (Archived by WebCite® at http://www.webcitation.org/63YmAWm9N)
- Achterman (2005). Surviving Wikipedia - improving student search habits through information literacy and teacher collaboration. *Knowledge Quest* **33**(5), 38-40.
- Alexandersson, M. & Limberg, L. (2003). Constructing meaning through information artefacts. *New Review of Information Behaviour Research* **4**(1),17–30.
- Boscolo, P., Ariasi, N., Favero, L. & Ballarin, C. (2011). Interest in an expository text: How does it flow from reading to writing? *Learning and Instruction* **21**(3), 467-480.
- Chu, S., Chow, K., Tse, S. & Kuhlthau, C. (2008). Grade 4 students' development of research skills through inquiry-based learning projects. *School Libraries Worldwide* **14**(1), 10-37.
- Davis-Lenski, S. & Johns, J.L. (1997). Patterns of reading-to-write. *Reading Research and Instruction* **37**(1), 15-38.
- Forte, A. & Bruckman, A. (2007). Constructing text: Wiki as a toolkit for (collaborative?) learning. In *Proceedings of the 2007 international symposium on Wikis (WikiSym '07). Oct 21-25, 2007, Montreal, Quebec, Canada.* (pp. 31-42). New York, NY: ACM Press.
- Forte, A. & Bruckman, A. (2010). Writing, citing, and participatory media: wikis as learning environments in the high school classroom. *International Journal of Learning and Media* **1**(4), 23-44.
- Gross, M. (2005). The imposed query. In K. Fisher, S. Erdelez & L. McKechnie (Eds.) *Theories of information behavior*. (pp. 164-168). Medford, NJ: Information Today.
- Harouni, H. (2009). High school research and critical literacy: social studies with and despite Wikipedia. *Harvard Educational Review* **79**(3), 473-494.
- Head, A. & Eisenberg, M. (2010). How today's college students use Wikipedia for course related research. *First Monday* **15**(3-1). Retrieved 12 July 2011 from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2830/2476. (Archived by WebCite® at http://www.webcitation.org/63UVHSERw)
- Heinström, J. (2002). *Fast surfers, broad scanners and deep divers - personality and information seeking behaviour*. Åbo (Turku), Finland: Åbo Akademi University Press. Retrieved 27 November, 2011 from http://users.abo.fi/jheinstr/thesis_heinstrom.pdf (Archived by WebCite® at http://www.webcitation.org/63UY4fEqb)
- Heinström, J. (2006). Fast surfing for availability or deep diving into quality – motivation and information seeking among middle and high school students. *Information Research* **11**(4), paper 433. Retrieved 4 May 2011 from http://informationr.net/ir/11-4/paper265.html. (Archived by WebCite® at http://www.webcitation.org/63UYCA7sJ)
- Huvila, I. (2010). Where does the information come from? Information source use patterns in Wikipedia. *Information Research* **15**(3), paper 433. Retrieved 12 July 2011 from http://InformationR.net/ir/15-3/paper433.html. (Archived by WebCite® at http://www.webcitation.org/63UYG5Fxt)
- Ingwersen, P. & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval.* New York, NY: Springer-Verlag.
- Jennings, E. (2009).Using Wikipedia to teach information literacy. *College & Undergraduate Libraries* **15**(4), 432-437.

- Julien, H. & Baker, S. (2009). How high-school students find and evaluate scientific information: A basis for information literacy skills development. *Library & Information Science Research* **31**(1), 12–17.
- Kuhlthau, C.C. (2004). *Seeking meaning: a process approach to library and information services.* 2nd ed. Westport, CT: Libraries Unlimited.
- Li, D.D. & Lim, C.P. (2008). Scaffolding online historical inquiry tasks: a case study of two secondary school classrooms. *Computers & Education* **50**(4), 1394-1410.
- Lim, S. (2009). How and why do college students use Wikipedia? *Journal of the American Society for Information Science and Technology* **60**(11), 2189-2202.
- Limberg, L. (1999). Experiencing information seeking and learning. *Information Research* **5**(1) paper 68. Retrieved 12 July 2011 from http://informationr.net/ir/5-1/paper68.html. (Archived by WebCite® at http://www.webcitation.org/63UYQSoYP)
- Limberg, L., Alexandersson, M., Lantz-Andersson, A. & Folkesson, L. (2008). What matters? Shaping meaningful learning through teaching information literacy. *Libri* **58**(2), 82–91.
- Lyut, B. & Tan, D. (2010). Improving Wikipedia's credibility: references and citations in a sample of history articles. *Journal of the American Society for Information Science and Technology* **61**(4), 715-722.
- Mateos, M. & Solé, I. (2009). Synthesising information from various texts: a study of procedures and products at different educational levels. *European Journal of Psychology of Education* **24**(4), 435-451.
- McGinley, W. (1992). The role of reading and writing while composing from sources. *Reading Research Quarterly* **27**(3), 226-248.
- McGregor, J. & Streitenberger, D. (2004). Do scribes learn? Copying and information use. In M. K. Chelton and C. Cool (Eds.), *Youth information-seeking behavior: theories, models and issues* (pp. 95-118). Lanham, MD: Scarecrow Press.
- McGregor, J. & Williamson, K. (2005). Appropriate use information at the secondary school level: understanding and avoiding plagiarism. *Library and Information Science Research* **27**(4), 496-512.
- Nilsson, L. Eklöf, A. & Ottosson, T. (2005). *Copy-and-paste plagiarism: technology as a blind alley or a road to better learning.* Paper presented at the 33rd congress of the Nordic Educational Research Association (NERA) in Oslo, Norway, 10th to 12th of March. Retrieved 11 July 2011 from http://www.distans.hkr.se/illwebb/nfpf2005_copy_and_paste_final.pdf. (Archived by WebCite® at http://www.webcitation.org/63UYcmK64)
- Purdy, J.P. (2009). When the tenets of composition go public: A study of writing in Wikipedia. *College Composition and Communication* **61**(2), 351-373. Retrieved 12 July 2011 from http://datacenter2.aucegypt.edu/bgironda/rhet343/wikiwrite.pdf. (Archived by WebCite® at http://www.webcitation.org/63UYhcD7Z)
- Purdy, J.P. (2010). Wikipedia is good for you? In C. Lowe & P. Zemliansky (Eds.), *Writing spaces: readings on writings* **1**, 205-224. Retrieved 12 July 2011 from http://wac.colostate.edu/books/writingspaces1/purdy--wikipedia-is-good-for-you.pdf. (Archived by WebCite® at http://www.webcitation.org/63UYmAvlD)
- Raven, M. & Flanders, A. (1996). Using contextual inquiry to learn about your audiences. *Journal of Computer Documentation* **20**(1), 1-13.
- Segev-Miller, R. (2004). Writing from sources: The effect of explicit instruction on college students' processes and products. *Educational Studies in Language and Literature* **4**,5–33.
- Spivey, N.N. (1997). *The constructivist metaphor. Reading, writing, and the making of meaning.* San Diego, CA: Academic Press.

- Sundin, O. (2011). Janitors of knowledge: constructing knowledge in the everyday life of Wikipedia editors. *Journal of Documentation* **67**(5), 840-862.
- Tynjälä, P, Mason, L. & Lonka K. (Eds.) (2001). *Writing as a learning tool. Integrating theory and practice.* Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Williamson, K. & McGregor, J. (2006). Information use and secondary school students: a model for understanding plagiarism. *Information Research* **12**(1) paper 288. Retrieved 12 July 2011 from http://InformationR.net/ir/12-1/paper288.html. (Archived by WebCite® at http://www.webcitation.org/63UYwJxtf)
- Williamson, K., McGregor, J., Archibald, A. & Sullivan, J. (2007). Information seeking and use by secondary students: the link between good practice and the avoidance of plagiarism. *School Library Media Research* **10**, 25 pages. Retrieved 12 July 2011 from http://eric.ed.gov/PDFS/EJ851698.pdf. (Archived by WebCite® at http://www.webcitation.org/63UZ06RkI)

How to cite this paper

Sormunen, E. & Lehti&oum;, L. (2011). "Authoring Wikipedia articles as an information literacy assignment – copy-pasting or expressing new understanding in one’s own words?" *Information Research*, **16**(4) paper 503. [Available at http://InformationR.net/ir/16-4/paper503.html]